COS10022 PROJECT 3: INTRODUCTION TO DATA SCIENCE

HUY NGOC NGUYEN – 103582239

April 17, 2021

1. EXECUTIVE SUMMARY

This report is an analysis of car park occupation rates in Birmingham, UK, provided by the UCI Machine Learning Repository. The main purpose of this report is to build prediction models for two car parks with the same correlation in occupancy rate through the process of data analysis and data visualization. Based on the data entries of each occupancy in a specific time and date, a time series model would be the best fit to predict the occupancy rate. Specifically, the model used in this report is the ARIMA model. Some key findings can be found below here:

- Car parks were divided into 3 groups: BHM, 'Others', and Named
- The highest occupancy rate of a car park was RC01 with 77.66% while the lowest occupancy rate of a car park was NIA North with 8.01%.
- BHM occupancy rate had the highest average of 55.58%. 'Others' occupancy rate had an average of 42.75%. The Named occupancy rate had the lowest average of 35.95%.
- HL01 and BX01 car parks were chosen for correlation analysis. HL01 averaged 68.62% while BX01 averaged at 69.48%. Both graphs showed an increasing trend from November to December.
- Model building and preparation were conducted through KNIME Analytics. Each car park dataset was split into trained and test datasets, which would be used for the construction of ARIMA models.
- The ARIMA parameters were calculated using ACF and PACF models. The model's parameters were (13, 1, 0).
- The statistics of the ARIMA models showed an average RMSE of 15.4 and an average standard deviation of 14.6.
- Other factors such as location or parking price could potentially play a role in determining a car park's occupancy rate.

2. INTRODUCTION

In today's world, cities have the technology to track the number of cars entering and exiting their car parks. Not only is this technology used for monitoring purposes, but also it helps to determine the performance of each car park. If several car parks are always reaching occupancy on a specific date or if other car parks are not getting filled up on a specific month, then the cities need to understand the reasons behind the occupancy rate to further expand or demolish those parking lots. In this assignment, a sample of car parks and their occupancy is provided from Birmingham, UK. Accompanied by the date and time measurement, the question for this task is to predict the occupancy rate of these car parks based on those measurements. Overall, this data analysis will explore when and where these car parks are full or available by constructing a predictive model using Time Series analysis. Specifically, an ARIMA model will be provided in the report to visualize the prediction and the actuality of a car park's occupancy rate.

3. VISUALIZATION AND CORRELATION ANALYSIS

Given the dataset of 35,717 records of 30 car parks, the occupancy rate can be calculated by dividing the occupancy of a car park at the point of time from its capacity. For this report, the types of car parks can be divided into 3 sections: BHM, Others, and Named. BHM car parks contain any car park starting with BHM. 'Others' car parks contain car parks starting with 'Others'. Named car parks are the remaining car parks out of the two previous categories.

From the graph below (Figure 1), the average occupancy rate of all car parks is 48.66%, of which slightly half of the car parks have a higher rate than the average. The car park ending with RC01 has the highest occupancy rate of 77.66% while the car park labeled NIA North has the lowest occupancy rate of 8.01%. However, it is worth considering that the data points of RC01 only span from Dec 13 to Dec 18, and the data points of NIA North span from Oct 16 to Nov 30. Hence, these parking lots cannot be considered for any correlation analysis.

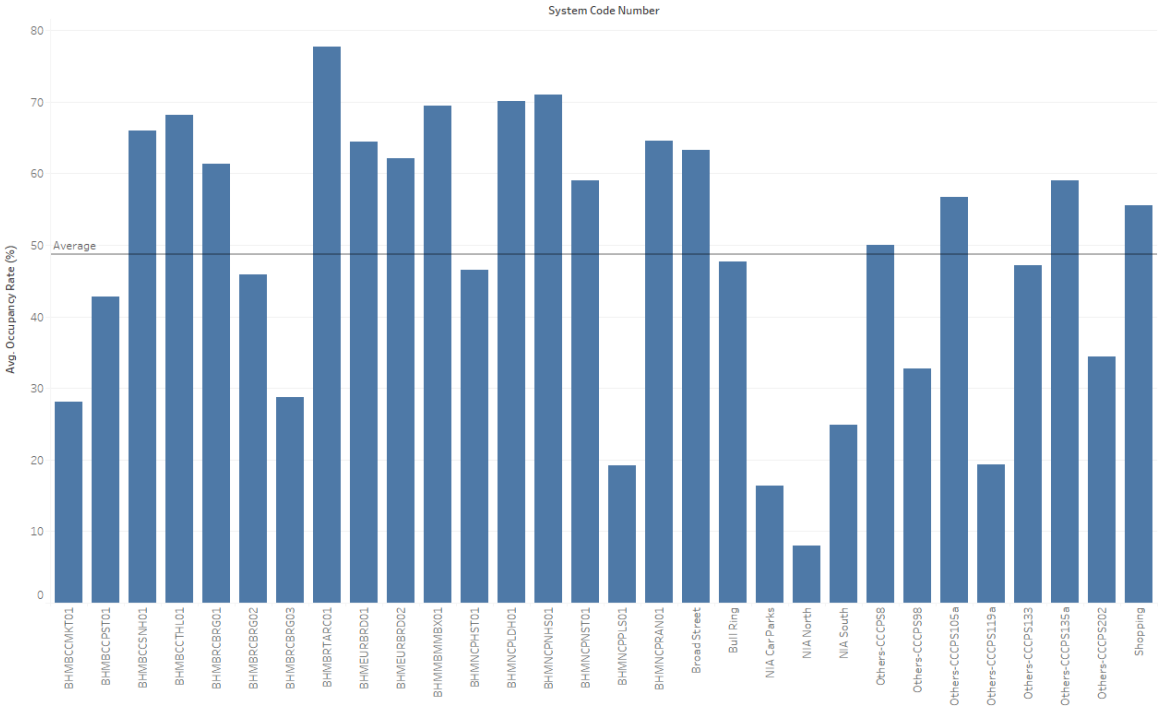

Figure 1: Occupancy rate of each car park with an average line

For BHM car parks, most parks record an increased occupancy rate during the weekdays and a plummeting rate during the weekends. On the other hand, 5 BHM car parks that have an initial lower average occupancy rate than the overall average produce the opposite trend: their occupancy rate increases during the weekends but falls off during the weekdays. The car park LS01 averages the lowest rate of 19.26% while the RC01 car park averages the highest rate of 77.66%. It is stated above that the RC01 graph only has a few data points, and therefore, if RC01 were to be omitted, the car park HS01 would have the highest occupancy rate of 71.04%. Except for RC01, every other car park can be considered for correlation analysis as their data span throughout the period (Figure 2).

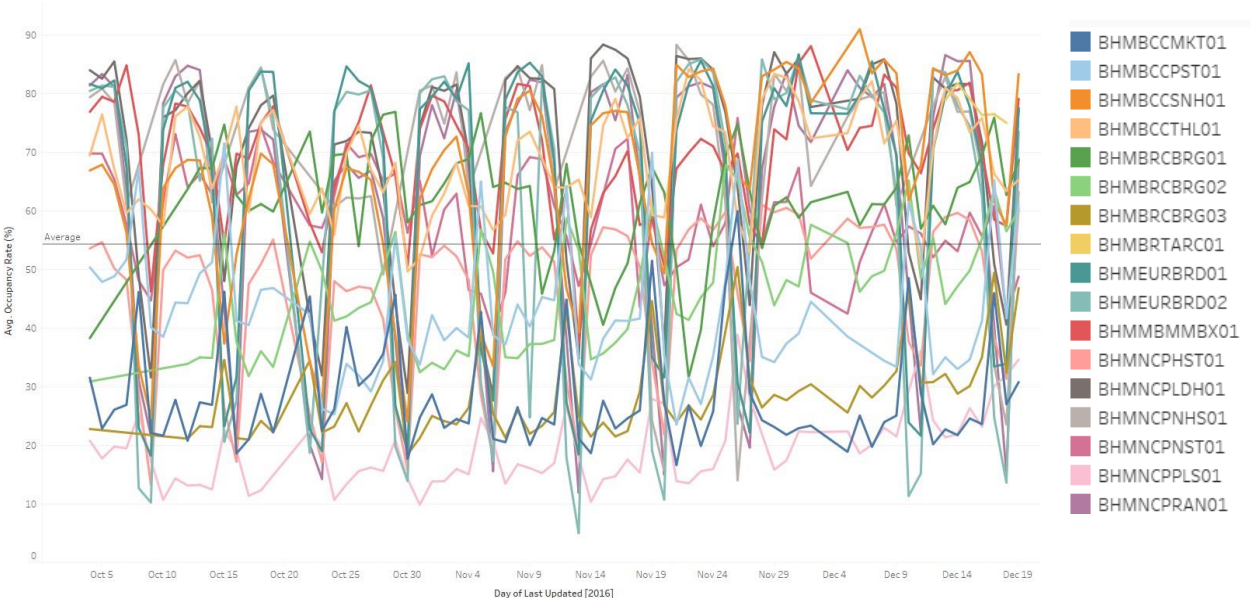

Figure 2: Occupancy rate of each BHM car park with an average line

For the 'Others' graphs, their data points all span from Oct 4 to Dec 19, which can be considered for correlation analysis. The 119a graph exhibits the lowest rate with an average of 19.27%, while the 135a graph displays a constant drastic change through each week, with the highest rate being 78.43% and the lowest rate being 18.92%, and impacting its average occupancy rate at only 59.06%. Unlike BHM graphs, the occupancy rate of every "Others" graph increases during the weekdays and decreases considerably on the weekends. Figure 3 shows that 3 out of 7 car parks have lower averages than the overall average occupancy rate of every car park.

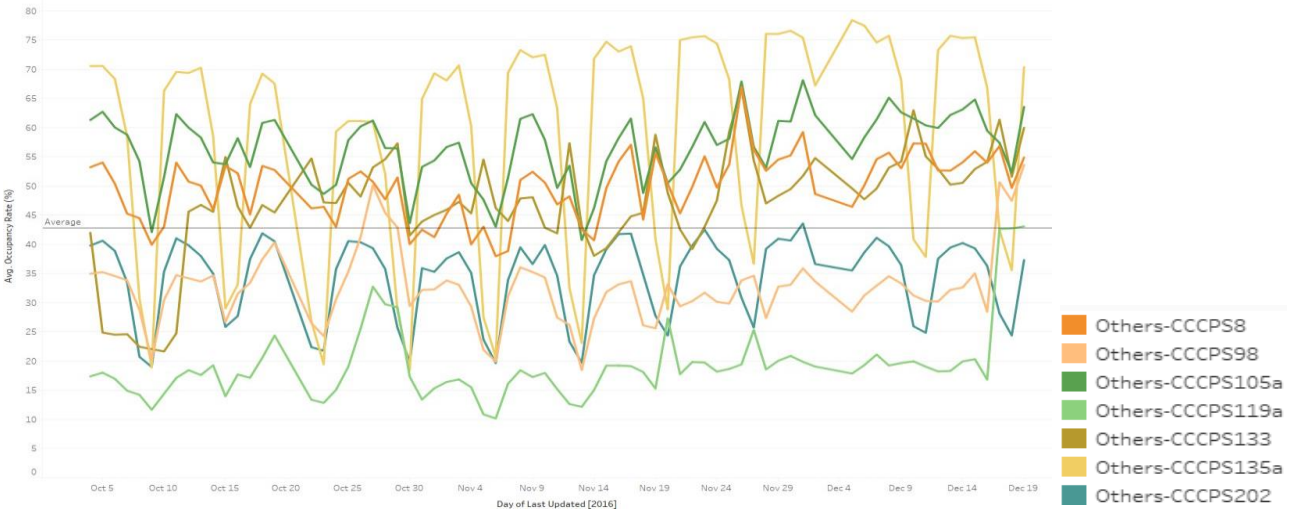

Figure 3: Occupancy rate of each 'Other' car park with an average line

For the Named car parks, NIA North exhibits the lowest occupancy rate and cannot be chosen for a correlation analysis due to a lack of data in other periods. Other NIA graphs with a lower average occupancy rate than the overall average of the Named car parks also have missing values from Dec 16 to Dec 19. Thus, those two graphs should not be examined further for correlation analysis. The three other graphs all have a higher average occupancy rate than the overall average, which is 41.28%. The "Broad Street" car park records its highest occupancy rate of 86.59% and its lowest occupancy rate of 14.18%, making its average occupancy rate to be 63.24%, which is the highest of every other Named car park (Figure 4).

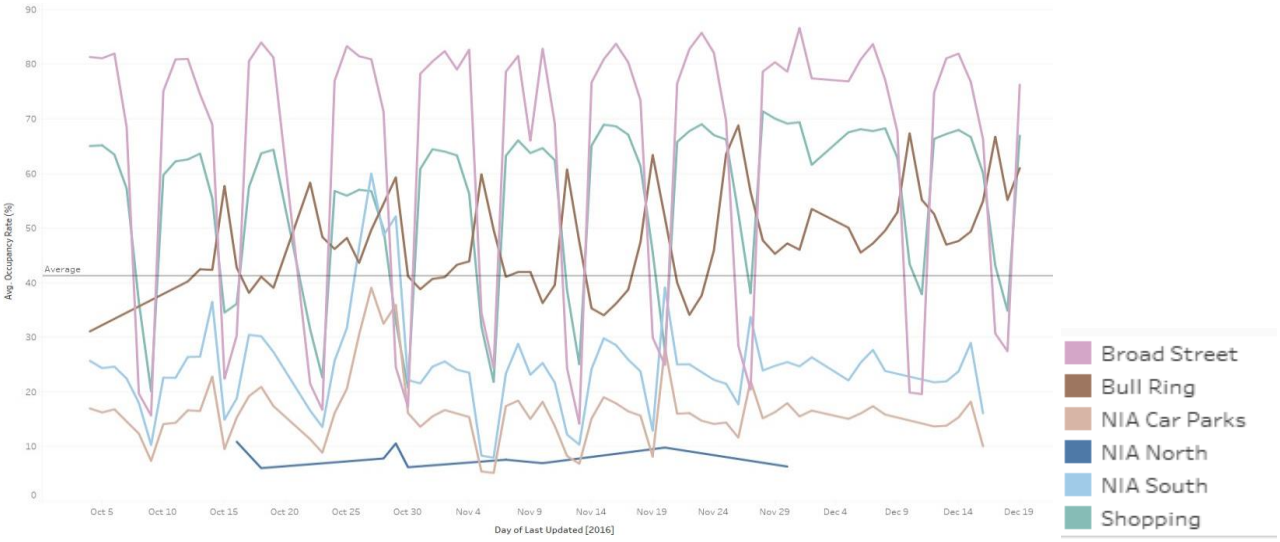

Figure 4: Occupancy rate of each "Named" car park with an average line

From the occupancy rate of each car park above, two car parks have been chosen to build prediction models: BHMBCCTHL01 and BHMMBMMBX01, which are shortened as HL01 and BX01 for easier reading. The figure below demonstrates the two car parks' occupancy rates for each day in the given time frame (Figure 5). Although BX01 (orange line) starts with a higher rate than HL01 (blue line), their shifts throughout each week are particularly similar. It is also worth noting that their averages are very close to one another, with the average of HL01 being 68.62% and the average of BX01 being 69.48%. Furthermore, when viewing the occupancy rate of both car parks in monthly periods, both are showing an increasing correlation towards December (Figure 6). Hence, these two car parks show a potential increasing trend in occupancy rates and will be used to build prediction models for future investments in expanding their capacity.

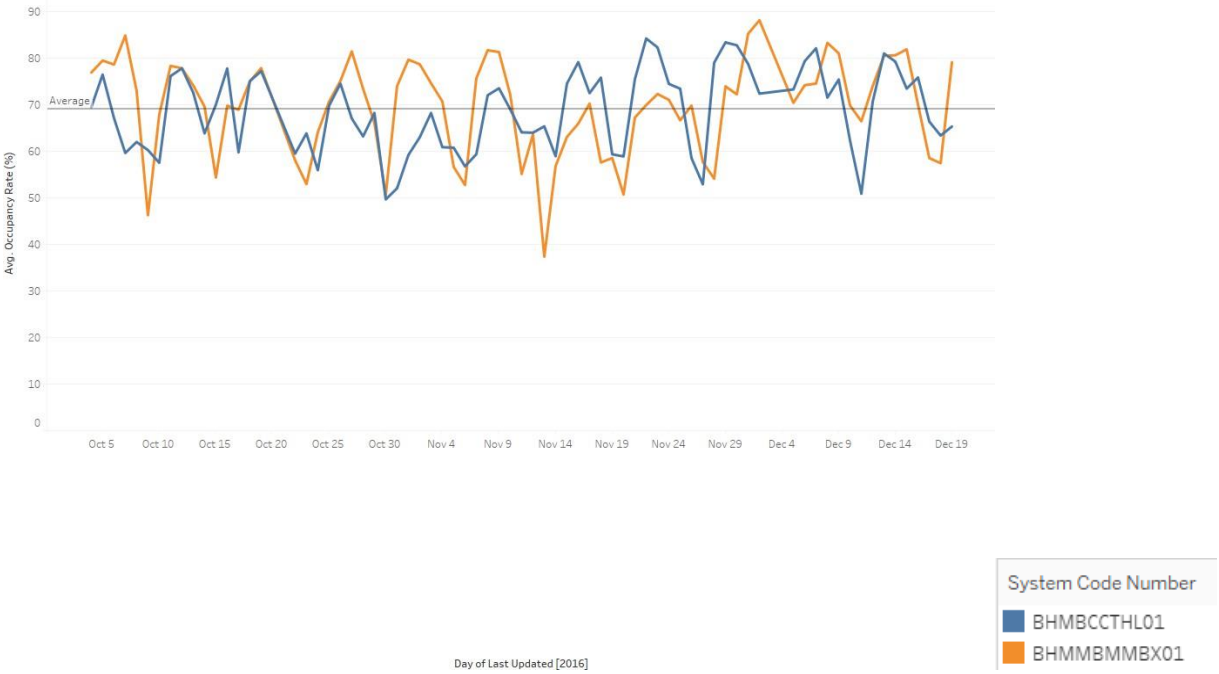

Figure 5: Occupancy rates of BX01 and HL01 through each day with an average line

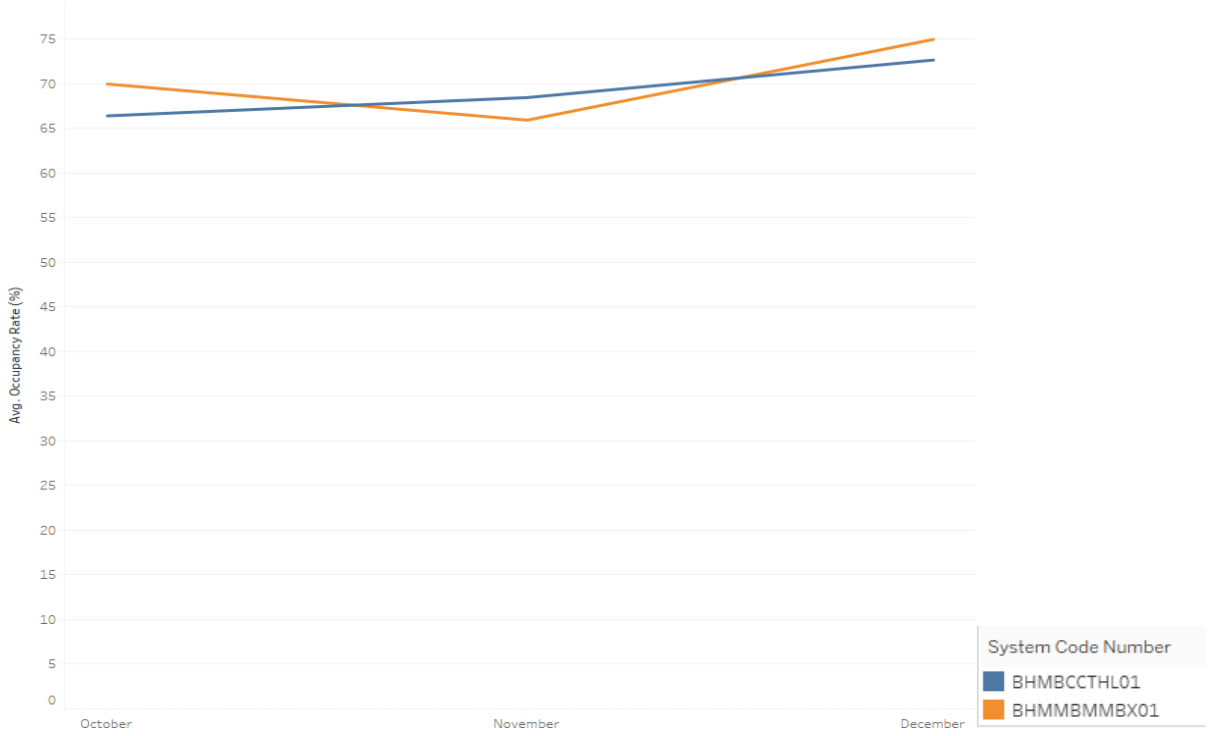

Figure 6: Occupancy rates of BX01 and HL01 through each month with an average line

4. MODEL BUILDING AND EVALUATION METRICS

Before constructing the model for the car parks, data cleaning is a necessary process to filter out any out-of-bounds data values. For this assignment, 385 data entries were deleted due to the occupancy rate excess, such as values less than 0% and more than 100%. Although there might be a few real-life cases where a car park occupancy exceeds its capacity, eliminating the outliers would ensure data consistency. After removing data errors, the dataset was imported into a KNIME workflow. The "LastUpdated" column containing the date and time for each data entry was then converted into date/time objects, which would be used later for the model. The Sorter node was used to arrange the date and time in ascending order, which would help avoid any data points from being placed in the wrong part of the model. The Column Filter node excluded the capacity and occupancy values from each data entry as they would be unnecessary variables to build the prediction model for these car parks.

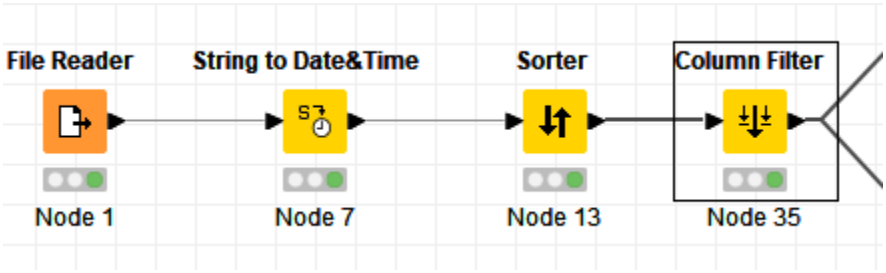

Figure 7: Importing and sorting workflow

Once the data was ready to be analyzed, the Row Filter node was used to filter out other car parks' data values, leaving only the values of HL01 and BX01 car parks. For this report, two models were built for each car park. For each model, their datasets were split into trained and test sets, with the trained set occupying 70% of the model data. Every partitioning method was completed through random sampling. For the model with data for 2 car parks combined, the dataset was concatenated and resorted using the "RowID" node, which would be suitable for use in the Partitioning node.

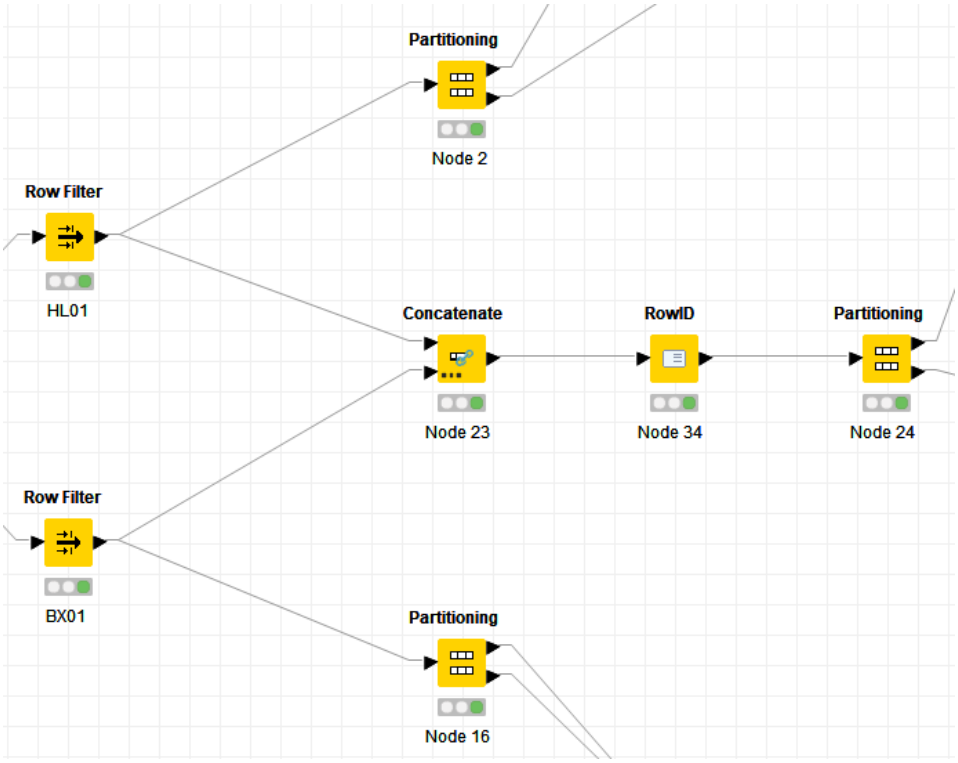

Figure 8: Filtering and partitioning nodes in the workflow

Since the objective of this report was to build predictive models for car parks based on their occupancy rate over each date and time, a time series model would be a perfect fit for this situation. Time-series graphs could assist in visualizing trends in numerical values over time, and because date and time are continuous categorical data, the data points generated along the x-axis can be connected by a single line. For this report, ARIMA models would be constructed to visualize the trained dataset of these car parks' occupancy rates. In addition, line plots would also be used to compare the predicted and actual results from the ARIMA model because combining the predicted model, the actual model, and the ARIMA model into one single graph is not possible in the KNIME platform. Other models such as k-means clustering, or random forest regression models are not suitable for this task because clustering and classification methods require more attributes of the dataset to be analyzed and grouped. Other time series models such as autoregression or moving average models could have been used in this report, but the utilization of the ARIMA model combines these two models with an order of differencing to help determine whether the predicted model has a stationary, a constant average, or a time-varying trend.

For the parameters of the ARIMA models, the visualization of the Autocorrelation Function (ACF) and the Partial ACF graphs are plotted below for lag observations. These graphs plot the correlation between the occupancy rate and its corresponding periods. The ACF graph displays a constant up and down pattern, and the graph peaks at the first lag and every other 13 increments (Figure 9). Moreover, because the original time series model demonstrates a constant average trend, the first order of difference will be used. The PACF graph shows a sharp cut-off at the lag-1 observation, and the data points are positively correlated for the following lag observations (Figure 10). Thus, the moving average degree would be zero for this baseline ARIMA model.

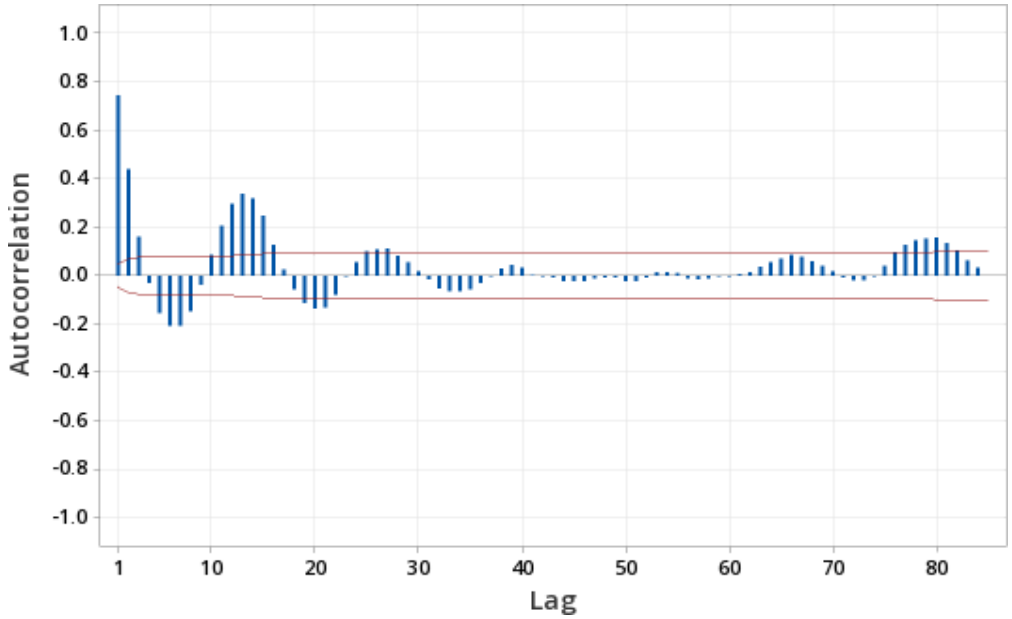

Figure 9: ACF graph of the original time series

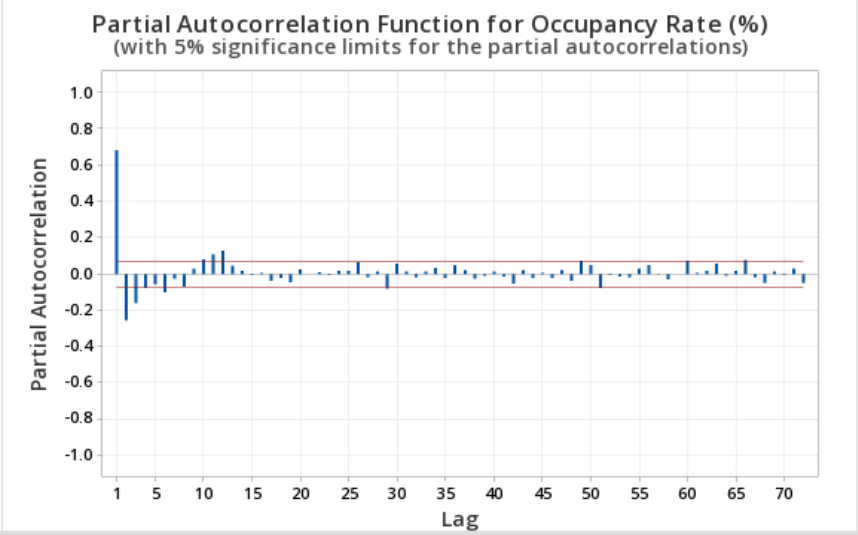

Figure 10: PACF graph of the original time series

Figure 11 indicates the workflow of the ARIMA model, in which the HL01 dataset is trained via the ARIMA Learner node and is tested via the ARIMA Predictor node. Then, the output of the ARIMA model is connected to three nodes: the Line Plot node for visualizations of the predicted and actual datasets, the Numeric Scorer node for computed statistics between the predicted and actual values, and the ARIMA Visualization node for the visualization of the ARIMA model.

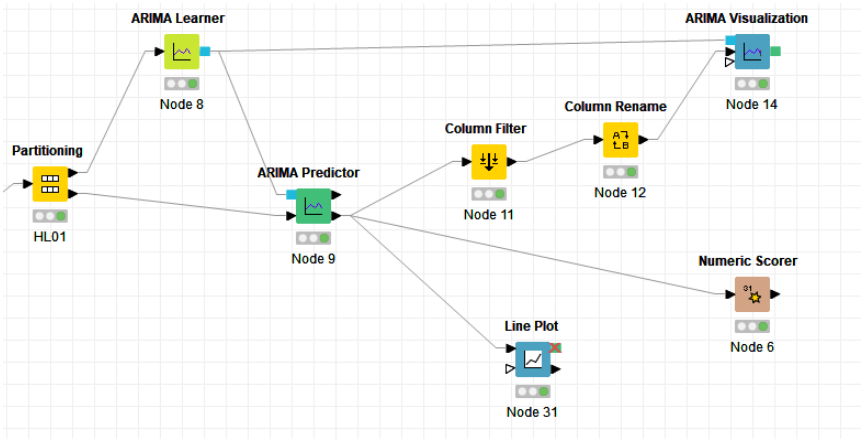

Figure 11: KNIME workflow for ARIMA model and its computational statistics

The graphs below are the predicted and actual values of the occupancy rate of the HL01 car park (Figure 12). Based on the shape of the line graphs, they do not entirely overlap with each other. The ARIMA model of the HL01 dataset shows a constant average trend across the timeline and forecasts the next several values, which reveals a similar trend with a narrower range. This can also be statistically validated through the RMSE and standard deviation measures.

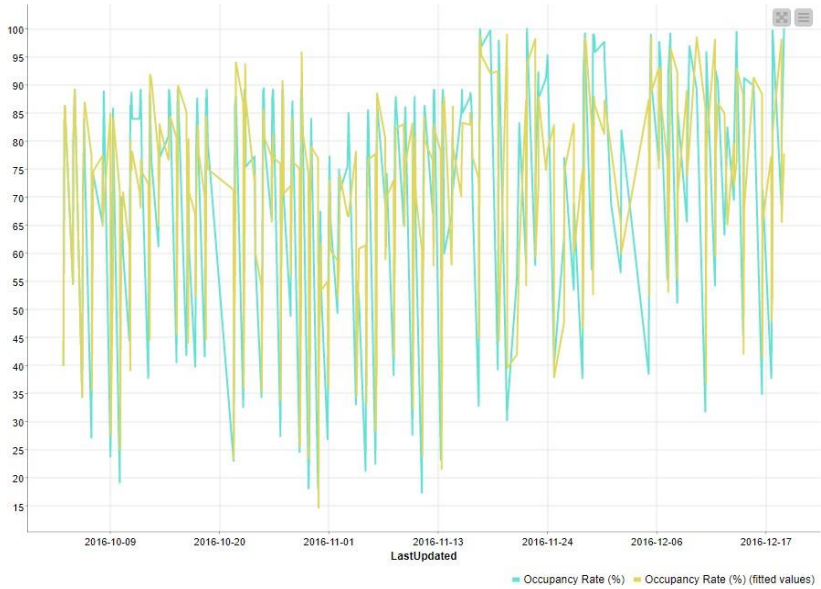

Figure 12: Actual (Blue) and Predicted (Yellow) results of the HL01 dataset

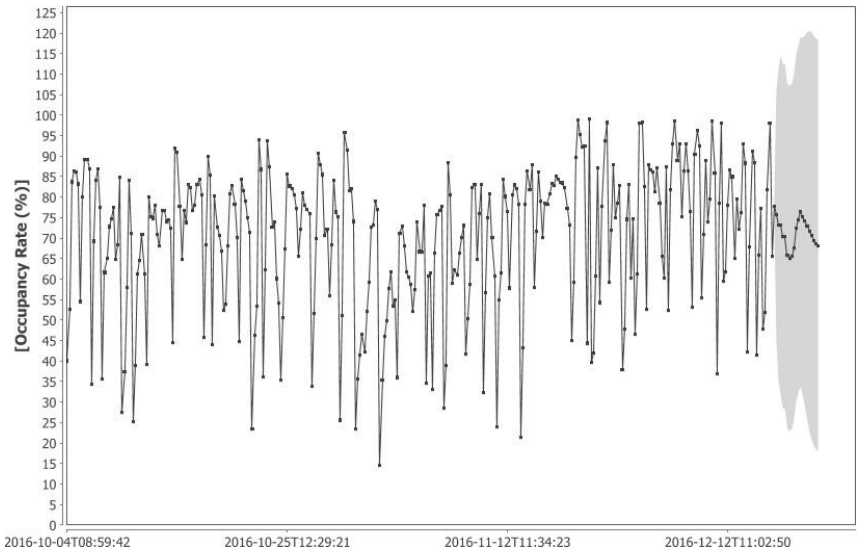

Figure 13: ARIMA model of the HL01 dataset

Similar to the HL01 comparable results, the graphs of the BX01 occupancy rate show a few non-overlapping sections. Most of the high extreme points on the graph are visualized by the predicted values, while most of the low extreme points are visualized by the actual values. The ARIMA model for the HL01 dataset demonstrates a constant average sideways trend, with the lowest data point averaging around 30% and the highest data point averaging higher than 100%. The forecasted values exhibit a similar correlation with the previous data points, but the range of the forecast is also narrower. The RMSE and standard deviation values of these models are slightly lower than the HL01 models.

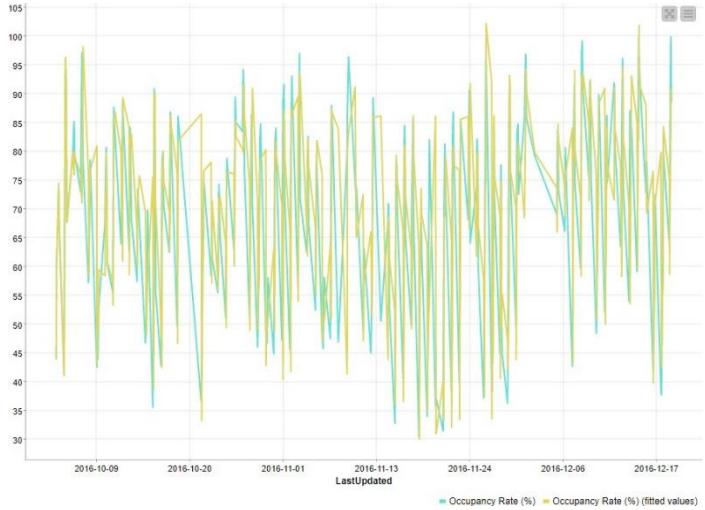

Figure 14: Actual (Blue) and Predicted (Yellow) results of the BX01 dataset

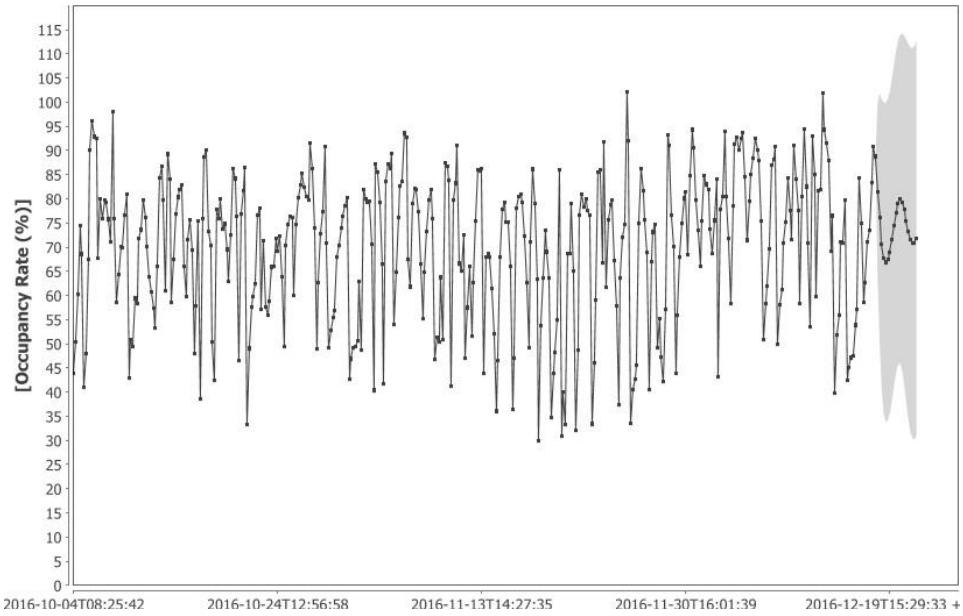

Figure 15: ARIMA model of the BX01 dataset

After building prediction models for the HL01 and BX01 car parks, the average occupancy rate of these datasets also increases moderately. The new average occupancy rate for the HL01 falls around 72%, which is 3% higher than its original average. Likewise, the new average occupancy rate for the BX01 is estimated at 73%, which is 4% higher than its original average.

5. POTENTIAL MISSING ATTRIBUTE

Assessing the occupancy rate of the car parks based only on the capacity and the occupancy of a specific date and time might be sufficient information to create a correlation analysis. However, other factors such as location and parking cost might add more value to understanding the occupancy rate of each car park. For instance, a city-centered car park might charge higher prices than a suburban car park due to a high volume of people moving in and out of the parking lot. Also, the price of a ticket might vary during different periods of the day, which impacts the number of occupancies in each period. Figure 16 is an example of a price and occupancy model from a San Francisco car park. If an owner wants to find the optimal occupancy rate for his car park, he or she must also consider the optimal parking price to keep up with customer demand as well as generate adequate revenue for future expansion.

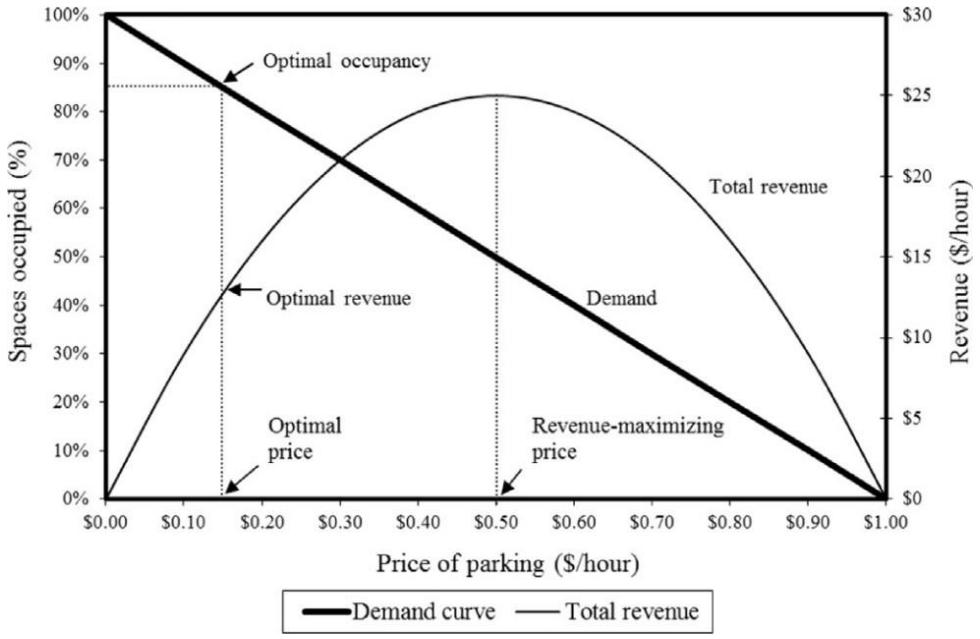

Figure 16: A model of a San Francisco car park with occupancy rate, parking price, revenue, and demand curve.

REFERENCES

Brownlee, J, 2017, 'How to Create an ARIMA Model for Time Series Forecasting in Python', *Machine Learning Mastery*, How to Create an ARIMA Model for Time Series Forecasting in Python (machinelearningmastery.com)

Chen, J, 2021, 'Autoregressive Integrated Moving Average (ARIMA)', *Investopedia*, *Autoregressive Integrated Moving Average (ARIMA) (investopedia.com)*

Duke University, 'Identifying the numbers of AR or MA terms in an ARIMA model', Identifying the orders of AR and MA terms in an ARIMA model (duke.edu)

Duke University, 'Introduction to ARIMA models', Introduction to ARIMA models (duke.edu)

Duke University, 'Rules for identifying ARIMA models', Rules for identifying ARIMA models (duke.edu)

Penn State University, 'Lesson 3: Identifying and Estimating ARIMA models; Using ARIMA models to forecast future values', Lesson 3: Identifying and Estimating ARIMA models; Using ARIMA models to forecast future values (psu.edu)

Pierce, G, Wilson, H, Shoup, D, 2015, 'Optimizing the use of public garages: Pricing parking by demand', *Transport Policy*, vol. 44, no. 89-95.